DOCUMENT RESUME

ED 178 555 TM 009 544

AUTHOR Benson, Jeri: Wilson, Michael
TITLE A Comparison of Three Types of Test Development Procedures Using Classical and Latent Trait Methods.
PUB DATE Apr 79
NOTE 27p.; Paper presented at the annual meeting of the American Educational Research Association (63rd, San Francisco, CA, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Statistics; Complexity Level; Efficiency; *Factor Analysis; *Item Analysis; Mathematical Models; Senior High Schools; Standard Error of Measurement; Statistical Analysis; *Test Construction; Test Items; Test Reliability; Test Selection
IDENTIFIERS *Item Discrimination (Tests); *Rasch Model

ABSTRACT

Three methods of item selection were used to select sets of 38 items from a 50-item verbal analogies test and the resulting item sets were compared for internal consistency, standard errors of measurement, item difficulty, biserial item-test correlations, and relative efficiency. Three groups of 1,500 cases each were used for item selection. First it was determined that, using a limit of 10 for mean square fit, there were 38 items that fit the Rasch model. Then the 38 items with the highest biserial item-test correlations were selected from the 50-item test, and the 38 items having the highest loadings on the first unrotated principal component were selected to make up the three item sets for comparison. There were 29 items common to all three sets. The 4,500 cases used for item selection were then randomly reassigned to three new samples to evaluate the three tests. The items selected by the Rasch method resulted in the most difficult and least internally consistent test, but all differences were small. Lord's procedures for approximating relative efficiency also showed the three tests to be quite similar over most of the score range. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178555

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Comparison of Three Types of Test Development Procedures Using Classical and Latent Trait Methods

Jeri Benson and Michael Wilson

University of Southern California

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JERI BENSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

TM009 544

Abstract

The present study was a follow up investigation of a comparative study (Benson, Crocker & Ware, 1978) involving three test development techniques: classical item analysis, factor analysis, and the Rasch model. The present investigation was conducted because in the original study a fixed number of items were used in comparing the three techniques. It may be that the arbitrarily selected number of items biased the Rasch technique. Therefore, three new tests were developed using the number of k items fitting the Rasch model as the criterion of item selection. The results indicated no substantive differences in terms of test precision and efficiency, thus implications for using the Rasch model solely as a test development technique were discussed.

A Comparison of Three Types of Test Development Procedures Using Classical and Latent Trait Methods

Since Binet and Simon (1916) first introduced the systematic approach to test development, psychometricians have been concerned with the extent to which accurate measurement of a person's "ability" is possible. The accuracy of the measurement or test precision is determined by the quality of the items in the test.

Classical item analysis is commonly used to identify single items which provide maximum discrimination between individuals on the latent trait being measured. Theoretically, selecting items which have high correlations with total test score will result in a discriminating test which is homogeneous with respect to the latent trait. Therefore, classical item analysis is an aid to developing internally consistent tests.

An alternative method of test development is factor analysis. Factor analysis is a statistical technique that takes into account the item correlation with all other individual items in the test simultaneously. Groups of similar items tend to cluster together and comprise the latent traits (factors) underlying the test. Classical item analysis then can be viewed as a unidimensional basis for test development, less sophisticated than the multidimensional procedure of factor analysis. However, classical item analysis and factor analysis share a basic weakness of being sample dependent. Critics of classical test theory contend that the item

statistics may vary when the examinee group changes. Item statistics may also vary if a different set of items from the same domain is used with the same examinee group (Hambleton & Cook, 1977; Wright, 1968). Thus, the selection of the final set of test items will be sample dependent.

Until recently, classical item analysis and factor analysis were the only techniques described in measurement texts for use in test development (Baker, 1977). However, since the publication of Lord and Novick's Statistical Theories of Mental Test Scores (1968) considerable attention has been directed toward the field of latent trait theory as a new area in test development. Proponents of latent trait theory claim that the advantages over classical test theory are twofold: (a) theoretically it provides item parameters which are invariant across examinee samples which will differ with respect to the latent trait, and (b) it provides item characteristic curves that give insight into how specific items discriminate among students of varying abilities (Lord & Novick, 1968; Wright, 1968).

Latent trait theory specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on a test (Hambleton & Cook, 1977). The relationship is described by a mathematical function; hence, latent trait models are mathematical models. Several latent trait models have been developed for use with dichtomously scored data: the normal ogive, and the one-, two-, and three-parameter logistic models. These models along with their mathematical formulations have been described by Hambleton and Cook (1977)

and Lord and Novick (1968). All four models are based on the assumption that the items in the test are measuring one common ability and the assumption of local independence exists between the items and the examinees. These two assumptions imply that a test which measures only one trait or ability will have less measurement error in the test score than a test that is multidimensional, and that the response of an examinee to one item is not related to his response to any other item. Where the latent trait models begin to differ is with respect to the shape of their item characteristic curves.

The normal ogive, developed by Lord (1952, 1953), and the two-parameter logistic model, developed by Birnbaum (1968) are essentially the same function (Baker, 1961), each providing an estimate of the item difficulty and discrimination parameters. The three-parameter model also developed by Birnbaum (1968), is an extension of the two-parameter model where the additional parameter being estimated is an index for guessing. The one-parameter logistic model, developed by Rasch (1960), is commonly referred to as the Rasch model. In addition to the two assumptions of latent trait models mentioned above, the Rasch model makes a third assumption, that all item discriminations are equal. Thus, the only parameter to be estimated using the Rasch model is item difficulty.

Of the various latent trait models that have been developed Wright (1977) has suggested that the one-parameter logistic Rasch model superior to other latent trait test development procedures due to the simplicity of the model. That is, the unweighted number right scoring

technique used by the Rasch model contains all the information necessary to produce estimators for the item and person parameters.

Benson, Crocker & Ware (1978) compared the Rasch model with classical and factor analytic methods of test development. The results of their study indicated no differences in the three test development techniques in terms of precision, but test efficiency levels were markedly different. The two classical methods produced tests that were more efficient for very low and very high ability students, and the Rasch method produced a test that was more efficient for average to high average ability students. A shortcoming of the investigation was that a fixed number of items were used in comparing the three techniques. It may well be that the arbitrarily selected 15 and 30 item tests reported in the Benson et al. study was biased against the Rasch technique in that too few or too many items were used that did not maximally fit the Rasch model.

The purpose of the present study was to conduct a follow-up investigation where the criterion of item selection was based on the Rasch model. Three tests were developed using the above criterion and compared in terms of test precision and efficiency. Sample size was not a variable in the current investigation since it was not found to be a significant variable in the previous study.

Method

An empirical study was designed to compare the effects of three test developmert procedures on a sample of high school students. The tests developed were compared on measures of:

* internal consistency
* standard error of measurement
* item difficulty
* item discrimination
* test efficiency curves

The study was designed in three phases: (a) item selection, (b) computation of item and test statistics for selected items on cross validation samples, and (c) statistical analyses of the test and item characteristics.

Subjects

A population of 78,751 high school seniors was available from 435 high schools throughout the state of Florida. From this population a 1 in 17 systematic sample of 4,500 students was chosen (Mendenhall, Ott & Scheaffer, 1971). The 4500 subjects were then randomly assigned to the three test development techniques, e.g. 1500 subjects in each group.

Instrument

The instrument used in this study was the Verbal Aptitude subtest of the Florida Twelfth Grade Test, a statewide assessment battery developed by the Educational Testing Service. The Verbal Aptitude subtest is comprised of 50 verbal analogies in a multiple choice format.

This particular type of instrument was selected for several reasons. First, as a cognitive measure of verbal ability it is similar to and representative of other national aptitude tests used for college admissions. Second, the subtest has been shown to be unidimensional (Benson et al., 1978) and thus the assumption of unidimensionality for the Rasch model is met. Third, other comparative latent trait studies have been conducted with similar item pools (Hambleton & Traub, 1973; Lord, 1968; Wright, 1968; Baker, 1961). Fourth, because of the length of time required to take such examinations the possibility of using fewer items while maintaining precision is desirable. Therefore, the question of which test development procedure would maximize the precision and thus the effectiveness of the test is an important one.

The Procedure

The data were analyzed in three phases. First, the selection of the items best fitting the Rasch model was based upon the mean square fit of the items to the model. The BICAL program (Wright & Mead, 1979) was used for the analysis. A mean square fit of 2.00 or less was arbitrarily

chosen as the criterion. The actual number of items determined by the above criterion were in turn used in the classical item analysis and factor analytic techniques. Under classical item analysis, the best items were selected based on the highest item biserial correlation with total test score. For factor analysis, the items were selected based on the highest item loadings on the first unrotated principal component. Second, the three test development techniques were reassigned randomly using a cross-validation design to a different sample of subjects. Then, the items selected by each method were submitted to a common item analysis program, GITAP (Baker & Martin, 1970) from which four of the five dependent variables in the study were obtained: (a) indices of internal consistency as measured by the analysis of variance procedure (Hoyt, 1941), (b) the standard error of measurement as defined by Magnusson (1966, p.79), (c) item difficulty, and (d) item biserial correlations. Third, statistical comparisons were made for three of the dependent variables: internal consistency estimates, item difficulty, and item discrimination. The observed internal consistency estimated from the test produced by the three techniques were compared to the projected population internal consistency estimate for a test of similar length using confidence intervals developed by Feldt (1965). Separate one-way analyses of variance were run for measures of item difficulty and item discrimination, the independent variable being the test development technique. If significant differences were indicated, Tukey's HSD post hoc procedure was employed. The standard error of measurement was not tested for statistical significance.

The fifth dependent variable in the study, a measure of relative efficiency for each of the tests was compared using the ability estimates

and procedures described by Lord (1974a, 1974b).

Test efficiency is the degree to which two unidimensional tests are able to discriminate between examinees of similar ability levels. The formula for approximating relative efficiency is (Lord, 1974b, p. 248):

$$R.E.\ (y,x) = (n_y/n_x) \cdot x(n_x - x)\ f_x^2\ /\ y(n_y - y)\ f_y^2\ ,$$

where R.E. denotes the relative efficiency of y compared to x, n_y and n_x denote the number of items in the two tests, x and y are the number-right scores having the same percentile rank, and f_x^2 and f_y^2 are the squared observed frequencies of x and y. Lord has suggested that the above formula only be used with a large sample of examinees and tests that are not extremely short.

Results

The data were analyzed and reported in the following manner: item selection, cross-validation, and comparison of the tests on measures of precision and efficiency.

Item Selection

The criterion of item selection was based upon the mean square fit of the items to the Rasch model. Using a mean square fit of 2.00 or less as proposed in the procedure, only nine of the fifty items fit the Rasch model. It was decided that this very small number of items was not sufficient to get stable estimates of the dependent variables in the study. Since Lord (1974b) has noted that in making relative efficiency

comparisons the test length should not be too short, the arbitrary criterion of 2.00 or less was revised upward to include a sufficient number of items in order to obtain stable estimates of the item parameter, and test efficiency comparisons. Therefore, a mean square fit of less than 10.00 was selected as the criterion. This resulted in 38 of the 50 items being selected by the Rasch procedure. The items selected under classical item analysis were determined by the 38 items having the highest biserial correlations with total test score. The selection of items under factor analysis was determined by the 38 items having the highest item loadings on the first unrotated principal component.

The items selected by each test development technique have been summarized on Table 1.

Insert Table 1 About Here

As noted on Table 1, 29 of the 38 items were common to all three test development techniques (76 percent overlap).

Cross-Validation

Once the 38 items were selected, the three samples were randomly reassigned to a different sample in order to obtain four of the dependent variables in the study. The GITAP item analysis program was used to estimate: internal consistency, the standard error of measurement, item difficulty and biserial correlations. In addition, the test mean and

standard deviation was obtained.

The descriptive statistics based on the cross-validation of the three 38 item tests are presented in Table 2.

Insert Table 2 About Here

As noted in Table 2, the Rasch developed test appeared to select more difficult items in that the mean of the Rasch test was lower than the mean of the classical and factor analytic tests (19.71, 21.33, 21.77 respectively).

The internal consistency estimates were highly similar for each method. The internal consistency estimate of the classical and factor analytic methods was .86 for each and .83 for the Rasch method. The confidence intervals discussed in the procedure section were not calculated because the results were so similar that it would be merely an excess of precision. Thus the three measures were assumed equivalent. A finding similar to the one noted in Benson et al. (1978).

There were essentially no differences among the three standard errors of measurement for the separate tests, 2.86 for classical, 2.71 for factor analysis, and 2.66 for Rasch.

Statistical Analyses

Separate one-way analyses of variance (ANOVA) were run on item

difficulty and item discrimination.

The ANOVA for item difficulty produced no significant differences between the three tests at the .05 level. The means were .56, .57, .52 for the classical, factor analytic, and Rasch techniques respectively (See Table 2).

The ANOVA for item discrimination produced no significant differences between the three tests at the .05 level. The means were .53, .53, .48 for the classical, factor analytic and Rasch item selection techniques respectively (See Table 2).

Relative Efficiency Comparisons

Relative efficiency is a means of comparing the precision of a test, but unlike internal consistency which is an integral part of the classical model relative efficiency is not based on any one test development methodology (Lord, 1974a, 1974b).

Three relative efficiency comparisons were made using the three 38 item tests. The three relative efficiency comparisons were: (a) the test developed by factor analysis was compared to the test developed by Rasch analysis, (b) the test developed by classical item analysis was compared to the test developed by Rasch analysis, and (c) the test developed from factor analysis was compared to the test developed by classical item analysis.

The efficiency curves for the three comparisons were shown in Figure 1.

Insert Figure 1 About Here

The relative efficiency value was plotted on the ordinate, while the percentile rank (student ability level) was plotted along the abscissa. The student ability estimate was based on an unweighted number right scoring procedure. Since each test contained 38 items, a relative efficiency of 1.00 would indicate that the tests were equally efficient.

A careful examination of the relative efficiency formula on page 9 indicated that when $n_x = n_y$ and $x = y$, that it was the number of examinees at the specified ability level (f_x^2 and f_y^2) that determined the efficiency of the test. That is, the fewer examinees observed at a particular percentile rank, the better the test discriminates at that percentile rank. Therefore, test efficiency was equated with the level of discrimination the test was able to make between examinees, at various scores or percentile ranks.

The tests developed by classical methods (factor analysis and classical item analysis) were equally efficient for most ability levels. The only exceptions were at the 5th-10th percentile and the 99th percentile. The test based on factor analysis was considered more efficient at the low ability levels, while the test based on classical item analysis was more efficient at the high ability level.

The test developed by factor analysis was more efficient for the low scoring examinees (5th-10th percentile rank) when compared to the test developed from the Rasch procedures. Both tests were about equally

efficient for the average ability examinees, with the exception of the 55th and 85th percentile ranks where the test based on factor analysis was more efficient than the Rasch test. However, the Rasch test was found to be more efficient than the factor analysis test at the very high ability level (99th percentile rank).

The relative efficiency comparison between tests based on the classical item analysis and the Rasch analysis was centered very close to 1.0 for most ability levels, hence the two tests were considered equally efficient with the exception of 70th and 99th percentile ranks. At the 70th percentile rank the classical item analysis test was more efficient than the Rasch test, but at the 99th percentile the Rasch test was more efficient than the test based on classical item analysis.

Discussion and Conclusions

This study was conducted to determine which of the three test development procedures, i.e., classical item analysis, factor analysis, and the Rasch model, might produce the superior test in terms of precision and efficiency of measurement from a common item and examinee population. The Rasch model was used as the criterion to indicate the number of items to be selected for each test.

In terms of test precision, the data reported indicated no significant differences between the three procedures. And, in fact, the three methods tended to select the same items; 76 percent were common to all three methods. The relative efficiency produced little in the way of

thorough and conclusive superiority of one method over the other except at either ends of the ability continuum. This finding may be due to the small sample observed at the upper and lower ends of the ability continuum, and as Magnusson (1966, p. 26-28) has pointed out the fewer number of examinees at any one point the better the test is able to discriminate at that point. For example, the Rasch developed test was more efficient than either classical method at the 99th percentile rank. The number of examinees scoring at that level on the Rasch test was 9 compared to 16 examinees scoring at that level on the factor analysis test and 17 examinees scoring at that level on the classical item analysis test. Although these numbers are small relative to the sample size of 1500, the discrepancy was large enough to illustrate the concept of test efficiency. All other factors held constant, the fewer the number of examinees at a particular ability level the more efficient that test becomes in providing information about a student's true ability level. Benson et al., (1978) found a slight superiority in the Rasch developed test in terms of relative efficiency for average to high ability students when compared to classical methods. However, in terms of the present study no one test development procedure dominated the ability continuum as a superior method of test development except at the extremes of the continuum.

It appears from the results of this study and Benson et al. (1978) that all three methods tend to select the better items from the item pool as noted by the 76 percent overlap in items common to all three methods and no significant differences in terms of test item characteristics.

In comparing factor analysis, classical item analysis, and the Rasch model as test development techniques it is important to remember that the Rasch model was not developed solely as a means of item selection. The primary function or advantage of the Rasch model is to provide invariant item and ability parameters, thus overcoming many of the problems of classical test theory in the area of test construction and interpretation. The real benefits derived from Rasch developed tests may not be found in the selection of test items, but more from the invariant properties of the model. These properties include item statistics that do not vary when the item is given to a different set of examinees, and person ability estimates that do not vary when a different set of items are administered to a given examinee. These invariant properties have greatly aided in the development of tailored testing, e.g., tests developed for specific individuals based upon their ability level, and the study of information and relative efficiency functions. In other areas of measurement, the Rasch model has been extremely useful in test equating (Rentz & Bashaw, 1977) and in the study of item bias (Wright, Mead, & Draba, 1976). Thus, if practitioners are to use latent trait methodologies to aid in the development of their local tests the rationale should be based more on the invariant properties derived from the models rather than being viewed as a new and novel item selection technique.

In conclusion, it should also be noted that only one of several latent trait models, the Rasch model, was used in this comparative study of test development techniques. Perhaps future research with the

more technical two- and three-parameter logistic models would produce tests superior to those developed by classical test theory in terms of precision of measurement and relative efficiency. Before the findings of the present study can be generally accepted, replication is needed using not only other populations, but other instruments as well. It may be important to consider whether similar results would be obtained if another instrument were used whose items were not as well refined as the ones used in the present study.

Table 1

Descriptive Statistics for Items Selected by Each Test Development Procedure[a]

Item	Difficulty	Discrimination[b]	Selected Item[c]
1	.73	.33	C, F, R
2	.89	.39	C, F, R
3	.77	.31	C, F, R
4	.82	.37	C, F, R
5	.82	.43	C, F, R
6	.78	.25	C, F,
7	.60	.40	C, F
8	.73	.33	C, F, R
9	.67	.35	C, F, R
10	.74	.43	C, F
11	.68	.42	C, F, R
12	.63	.33	C, F, R
13	.52	.45	C, F
14	.79	.44	C, F, R
15	.63	.33	C, F, R
16	.67	.45	F, R
17	.51	.31	C, F, R
18	.66	.33	C, F, R
19	.70	.30	C, F
20	.52	.29	C, F, R
21	.50	.38	F
22	.52	.45	C, F, R
23	.49	.22	C, F, R
24	.46	.32	C, F, R
25	.53	.25	C, F, R
26	.53	.47	C, F, R
27	.36	.45	C, F, R
28	.58	.39	C, F, R
29	.45	.43	C, F, R
30	.46	.48	F
31	.56	.50	C, F, R
32	.44	.24	
33	.52	.41	
34	.38	.36	C, F, R
35	.42	.15	C, F, R
36	.37	.28	C, F, R
37	.35	.44	C, R
38	.35	.49	C, R
39	.35	.45	C, F, R
40	.35	.54	C, F, R

Table 1 cont.

Item	Difficulty	Discrimination[b]	Selected Item[c]
41	.28	.20	
42	.25	.20	R
43	.26	.21	R
44	.32	.55	C, R
45	.33	.21	F
46	.26	.26	
47	.27	.22	C, F, R
48	.22	.21	C, R
49	.16	.53	R
50	.18	.55	R

[a] 29 of the 38 items were common to all three tests (76 percent)

[b] as measured by point biserial correlations

[c] C = item selected by classical item analysis
F = item selected by factor analysis
R = item selected by Rasch analysis

Table 2

Descriptive Statistics for the 38 Item Tests Based on Each Item Analytic Procedure

N = 1500

	Classical Item Analysis	Factor Analysis	Rasch Model
Test:			
Mean	21.33	21.77	19.71
Standard Deviation	7.17	7.25	6.44
Internal Consistency[a]	.86	.86	.83
Standard Error of Measurement	2.68	2.71	2.66
Difficulty:			
Mean	.56	.57	.52
Standard Deviation	.18	.17	.20
Range	.24 - .90	.27 - .89	.18 - .91
Biserial:			
Mean	.53	.53	.48
Standard Deviation	.12	.13	.08
Range	.30 - .75	.26 - .75	.34 - .62

[a] as measured by Hoyt's analysis of variance procedure

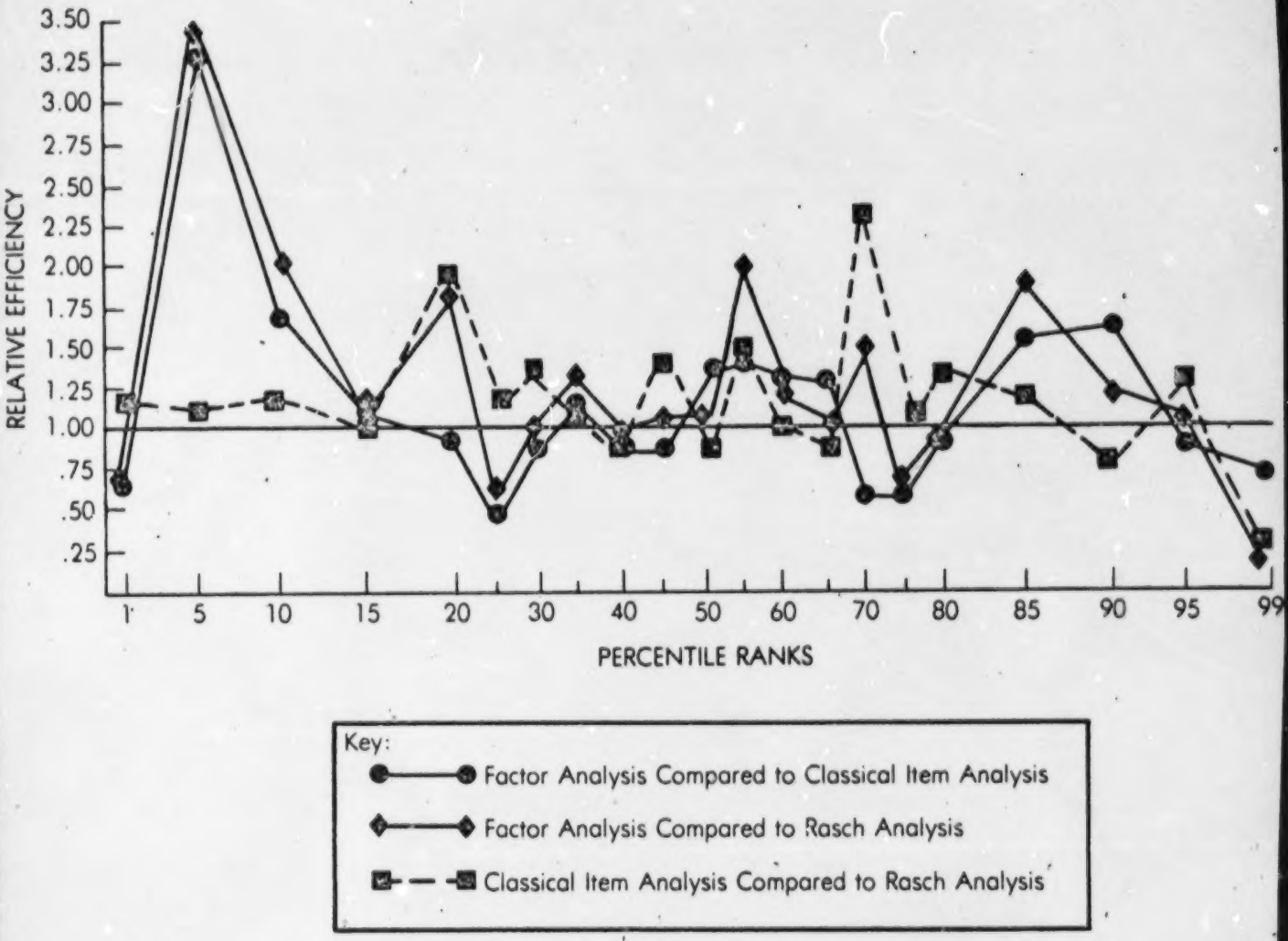

FIGURE 1. RELATIVE EFFICIENCY COMPARISONS FOR THE THREE 38-ITEM TESTS N = 1500

Appendix A
Relative Efficiency Values Used in Figure 1
for the Comparison Among Test Development Methods

Percentile Rank	Factor Analysis to Classical	Classical to Rasch	Factor Analysis to Rasch
1	.62	1.20	.70
5	3.20	1.10	3.40
10	1.70	1.20	2.00
15	1.10	1.00	1.20
20	.90	1.90	1.80
25	.50	1.20	.60
30	.80	1.30	1.00
35	1.20	1.10	1.30
40	.90	.90	.90
45	.90	1.40	1.10
50	1.30	.90	1.10
55	1.40	1.50	2.00
60	1.20	1.00	1.20
65	1.20	.80	1.00
70	.60	2.30	1.50
75	.60	1.10	.70
80	.90	1.30	.90
85	1.50	1.20	1.90
90	1.60	.80	1.20
95	.90	1.20	1.10
99	.80	.30	.20

REFERENCES

Baker, F.B. Empirical comparison of the item parameters based on the logistic and normal functions. Psychometrika, 1961, 26, 235-246.

Baker, F.B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Baker, F.B. & Martin, T. Fortap: A Fortan test analysis package. Occasional Paper No. 10. Office of Research Consultation, College of Education, Michigan State University, 1970.

Benson, J., Crocker, L. & Ware, W. A comparison of three types of item analysis in test development using classical and latent trait methods. A paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March, 1978.

Binet, A. & Simon, T. (The development of intelligence in children) (E.S. Kite, trans.). Baltimore, Md.: Williams & Wilkins, 1916.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Feldt, L.S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. Psychometrika, 1965, 30, 357-365.

Hambleton, R.K. & Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hambleton, R.K., & Cook, L. Latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 1977, 14 75-96.

Lord, F.M. A theory of test schores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548.

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M. The relative efficiency of two tests as a function of ability level. Psychometrika, 1974, 39, 351-358(a).

Lord, F.M. Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 1974, 11, 247-254(b).

Lord, F.M. & Novick, M. Statistical theories of mental test scores. Reading, Ma.: Addison-Wesley, 1968.

Magnusson, D. Test theory. Reading, Ma.: Addison-Wesley, 1966.

Mendenhall, W., Ott, L., & Schaffer, R. Elementary survey sampling. Belmont, Ca.: Wadsworth Publishing Co., 1971.

Rasch, G. Probablistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedagogiske Institute, 1960.

Rentz, R.R. & Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Woodcock, R.W. Woodcock Reading Mastery Test. Circle Pines, Minn.: American Guidance Service, 1974.

Wright, B. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. & Mead, R.J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B.D., Mead, R.J., & Draba, R. Detecting and correcting test item bias with a logistic response model. Research Memorandum No. 22, Statistical Laboratory, Department of Education, University of Chicago, 1976.